User Repurchase Prediction

Sean Matthews

github/sean-io/market-basket-analysis

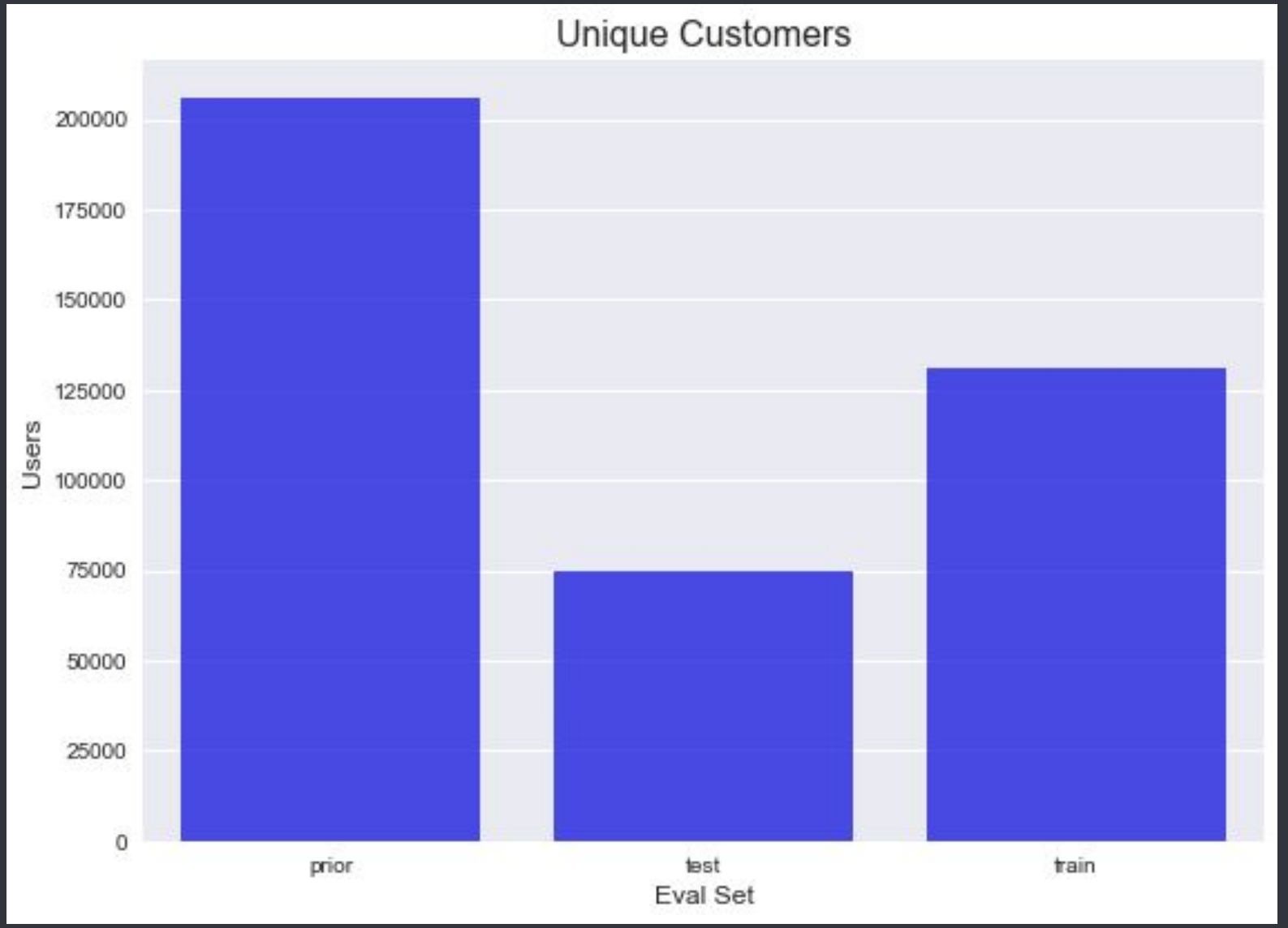
Objective & Context

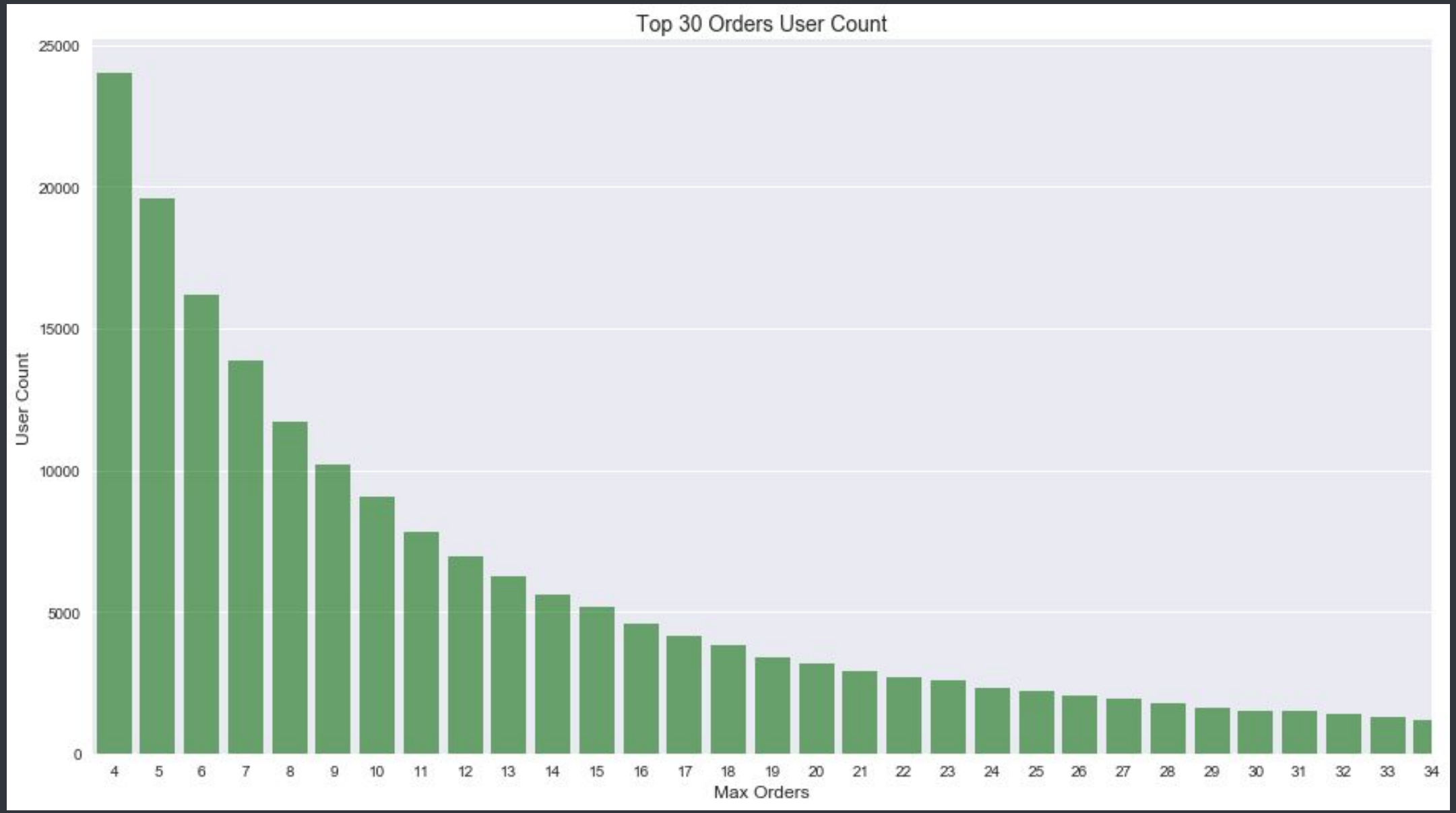
Predict which previously purchased products a user will order next.

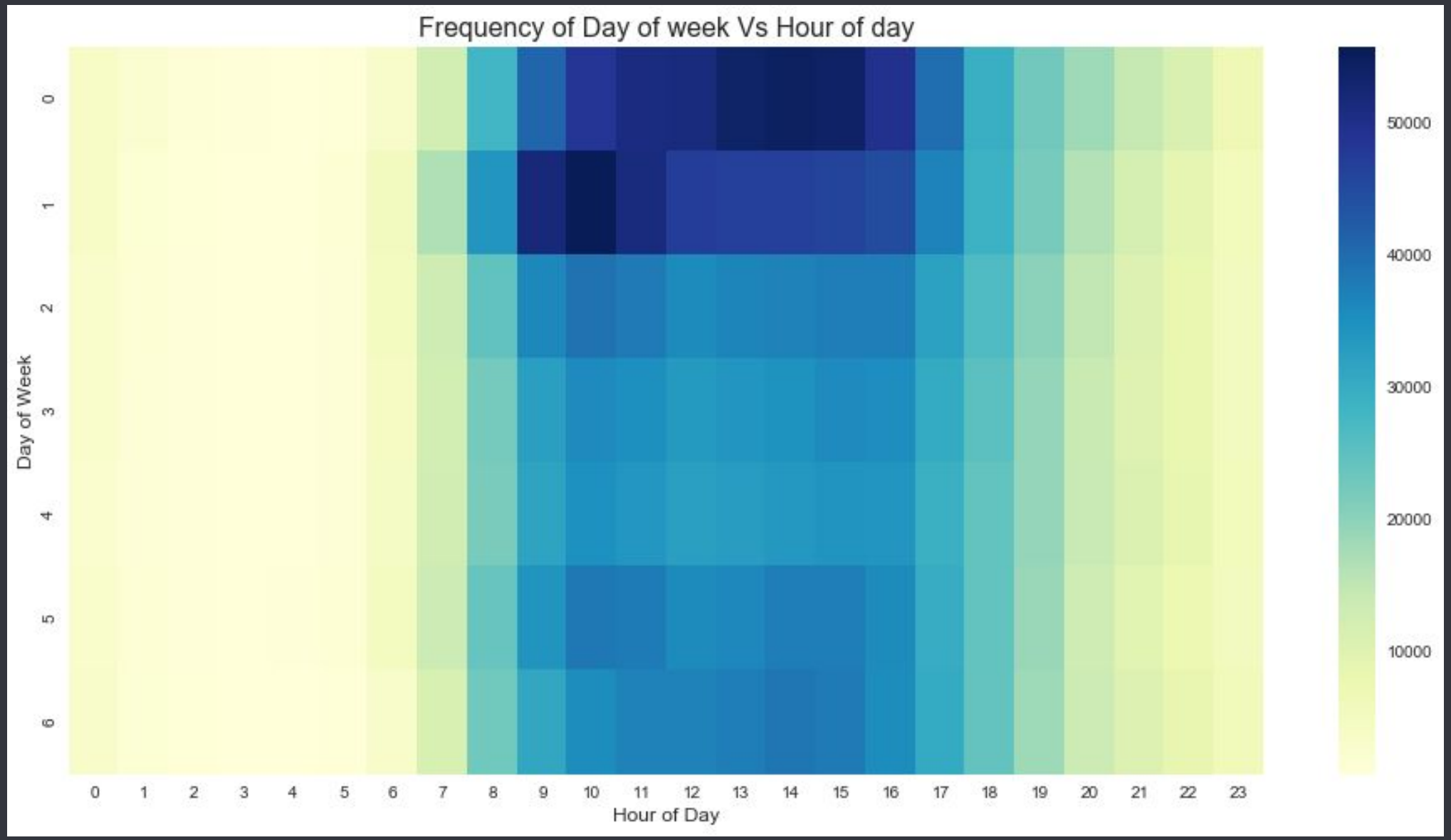
Kaggle Instacart Challenge Data

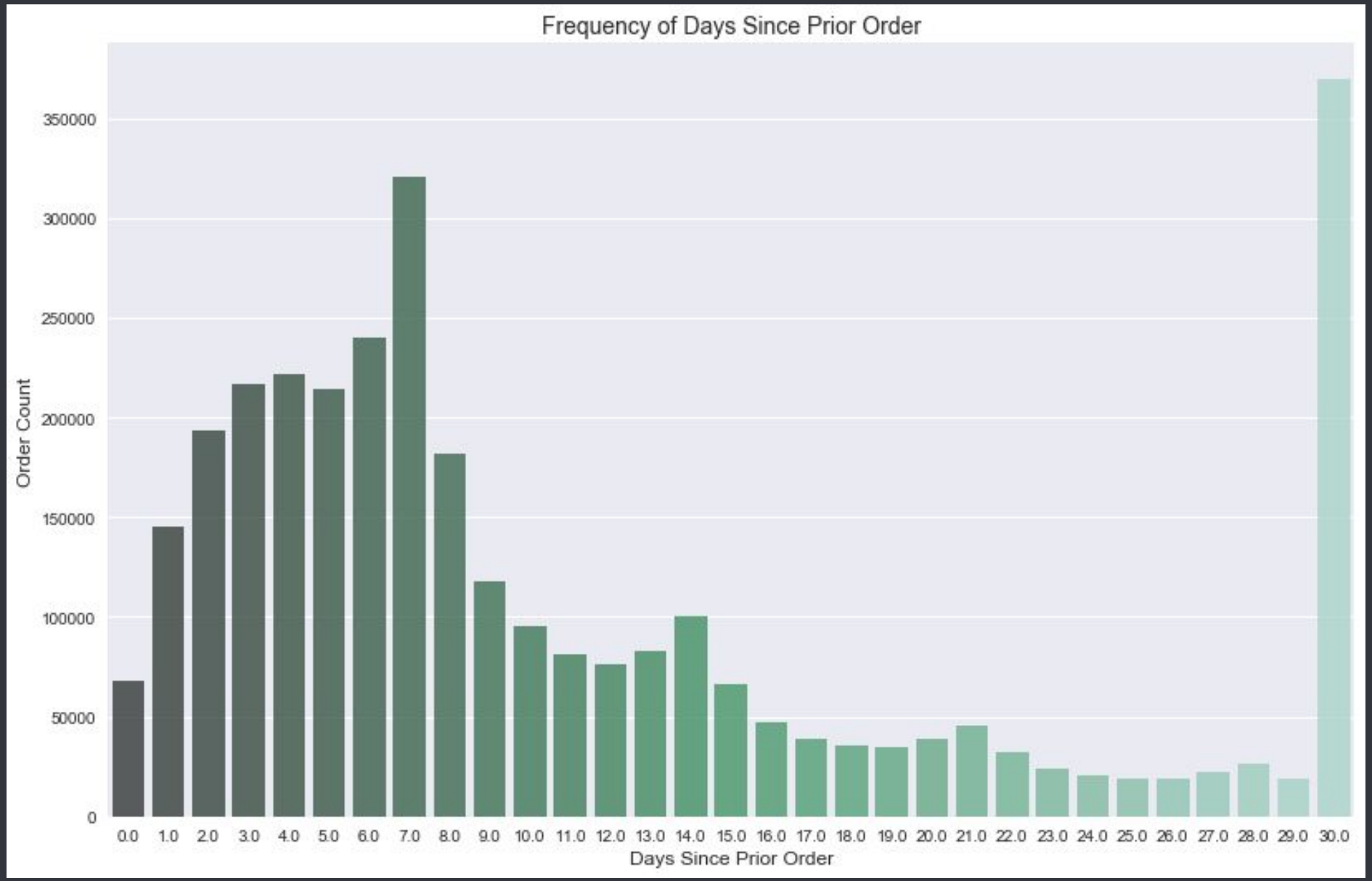
Data consists of over 3 million orders from more than 200K users.

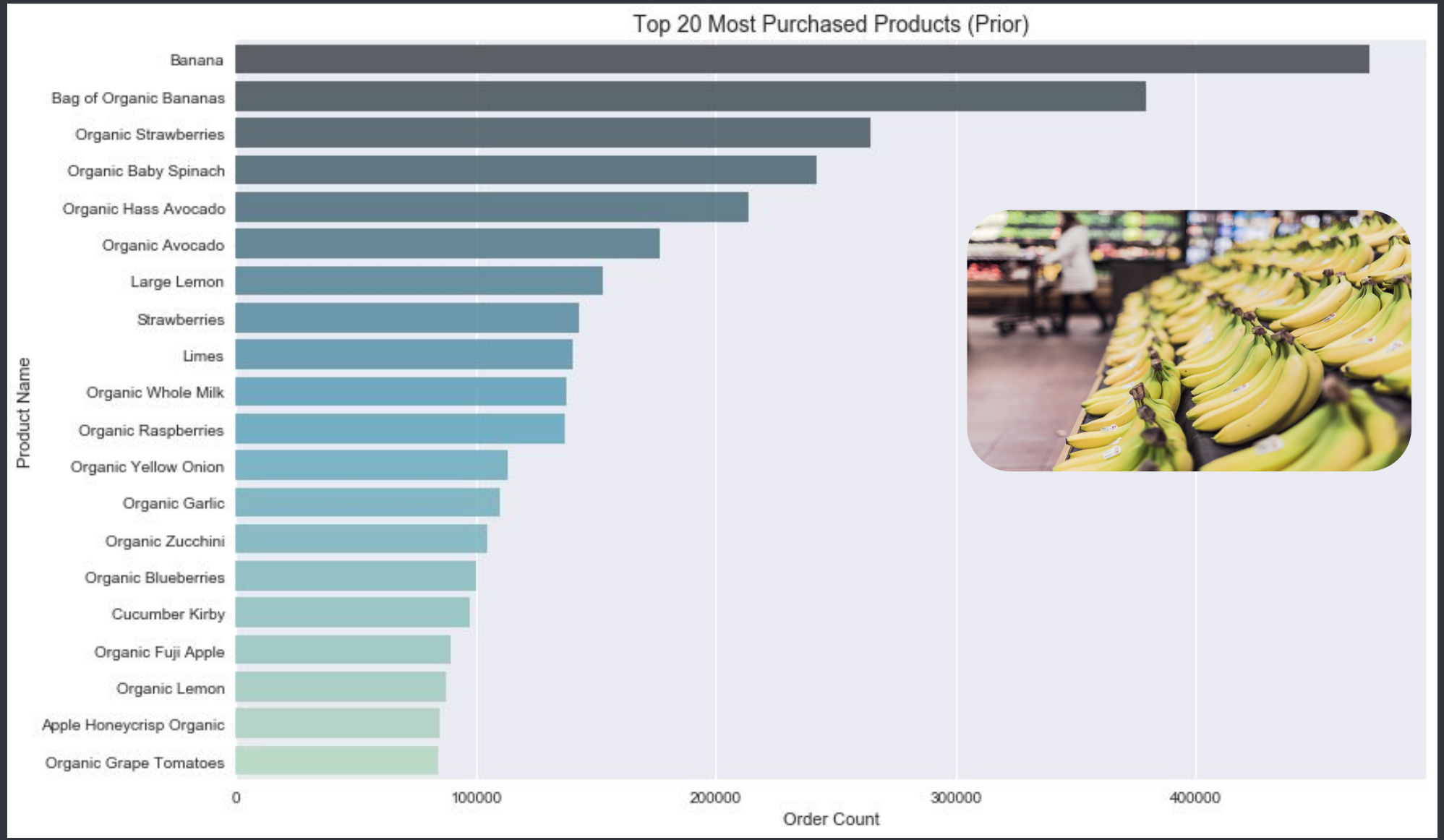
Available information: user order count, order products were added to a user's cart, day of week, hour of day, reordered indicator, product department, product aisle.


Unique Customers
Users
200000
175000
150000
125000
100000
75000
50000
25000
0
prior
test
train
Eval Set


Top 30 Orders User Count
User Count
25000
20000
15000
10000
5000
0
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Max Orders


Frequency of Day of week Vs Hour of day
Day of Week
0
1
2
3
4
5
6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day
50000
40000
30000
20000
10000


Frequency of Days Since Prior Order
Order Count
0
50000
100000
150000
200000
250000
300000
350000
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0 29.0 30.0
Days Since Prior Order


Top 20 Most Purchased Products (Prior)
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Apple Honeycrisp Organic
Organic Grape Tomatoes
Product Name
0
100000
200000
300000
400000
Order Count

Feature Selection

Product Features

Prod order count: product order count (by all users)

Prod reorders: product reorder count (by all users)

Prod reorder rate: frequency product is reordered (by all users)

User Features

User orders: user order count

Avg basket size: average number of products per order

Avg days between orders: average number of days between a user's orders

User Product Features

Avg add to cart order: average order user adds product to their cart (order)

User prod reorders: user reorder count of product

User prod reorder rate: user frequency of reordering product

Feature Selection

- ➔ Reduced training (prior order) data set to 1% of its original record count. 32M > 300K
- ➔ Utilized Random Forest Classifier to evaluate feature importance.
- ➔ ‘Avg days between orders’ was left out of features included in predictive models.

user_prod_reorders	0.295
user_prod_reorder_rate	0.279
prod_reorder_rate	0.048
user_orders	0.043
avg_days_between_orders	0.035

63.0%

Products reordered in the training (prior order) data set.

Products were reordered in the validation (test order) data set.

59.8%

Baseline performance

(Using SK Learn LR evaluation)

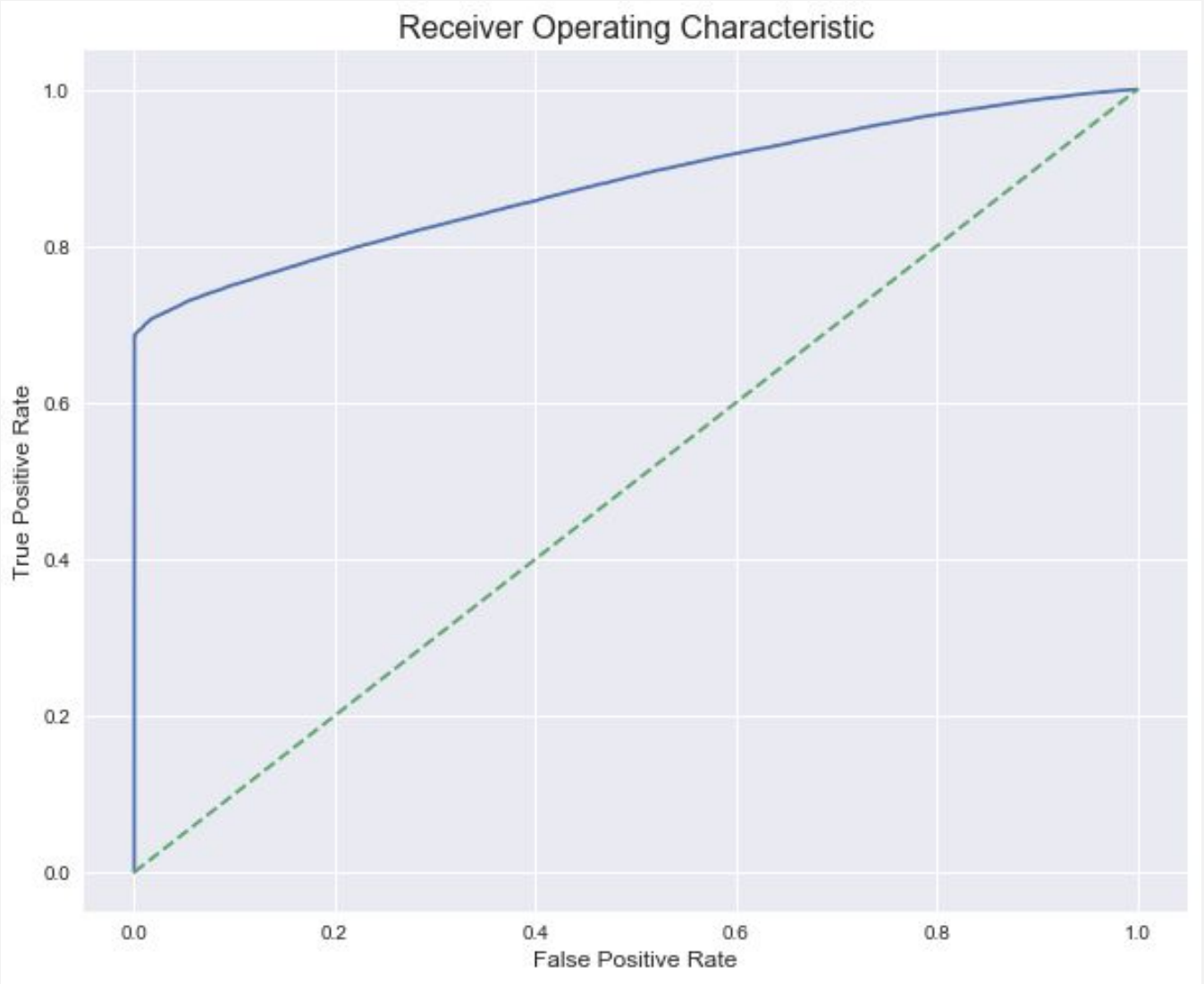
Accuracy 69.6%

Precision 100%

Recall 49.2%

F1 0.689

AUCROC 87.8%

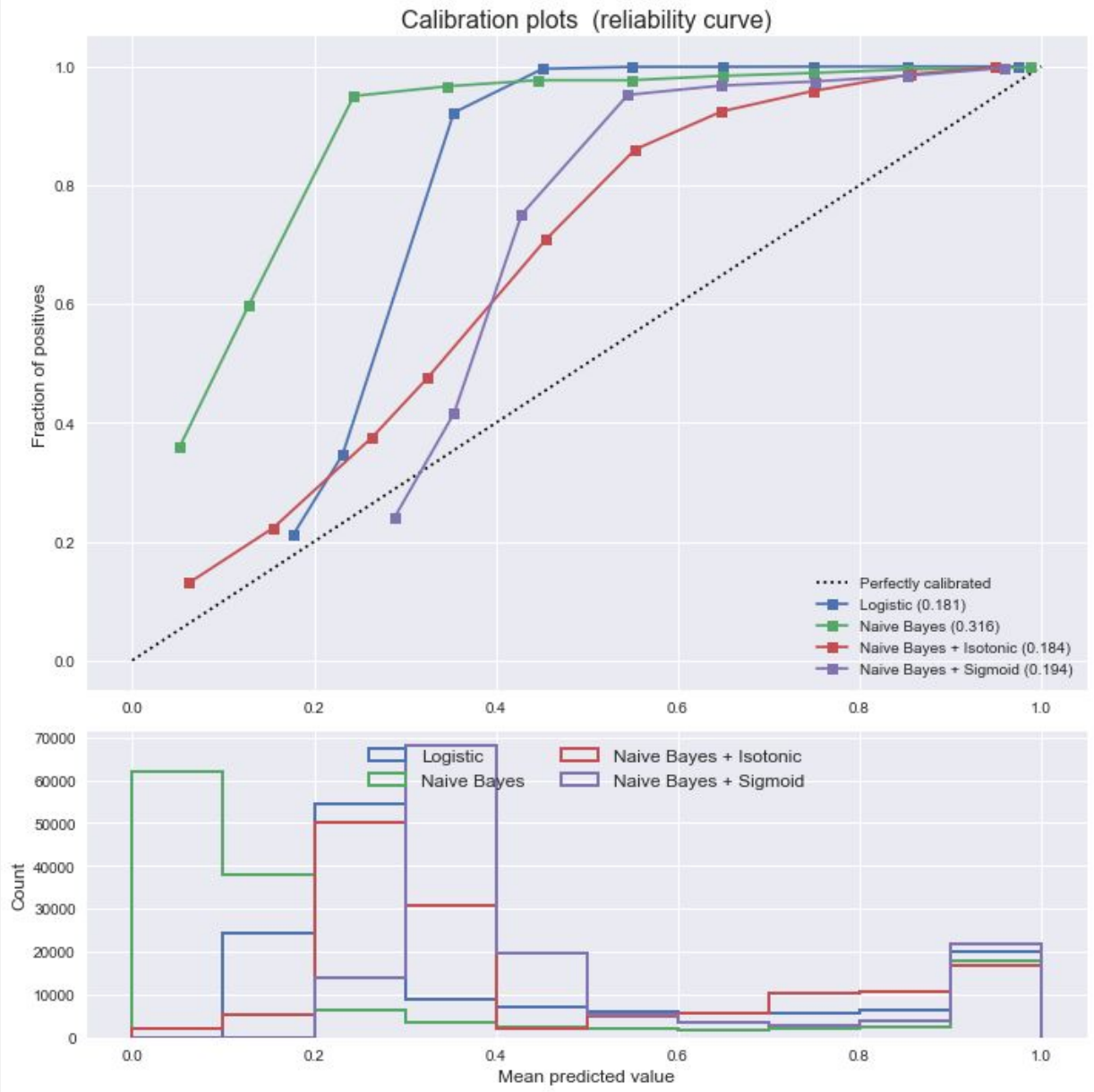


Other SK Learn Models' AUROC

88.6% GBDT

86.6% RF

86.3% KNN



SK Learn Calibration Plots

Logistic:
- Precision: 1.000
- Recall: 0.526
- F1: 0.689

Naive Bayes + Isotonic:
- Precision: 0.965
- Recall: 0.560
- F1: 0.709

Naive Bayes:
- Precision: 0.995
- Recall: 0.315
- F1: 0.478

Naive Bayes + Sigmoid:
- Precision: 0.985
- Recall: 0.439
- F1: 0.607

SK Learn Calibration Plots

Logistic:

Precision: 1.000

Recall: 0.526

F1: 0.689

SVC:

Precision: 0.665

Recall: 0.934

F1: 0.777

SVC + Isotonic:

Precision: 0.994

Recall: 0.636

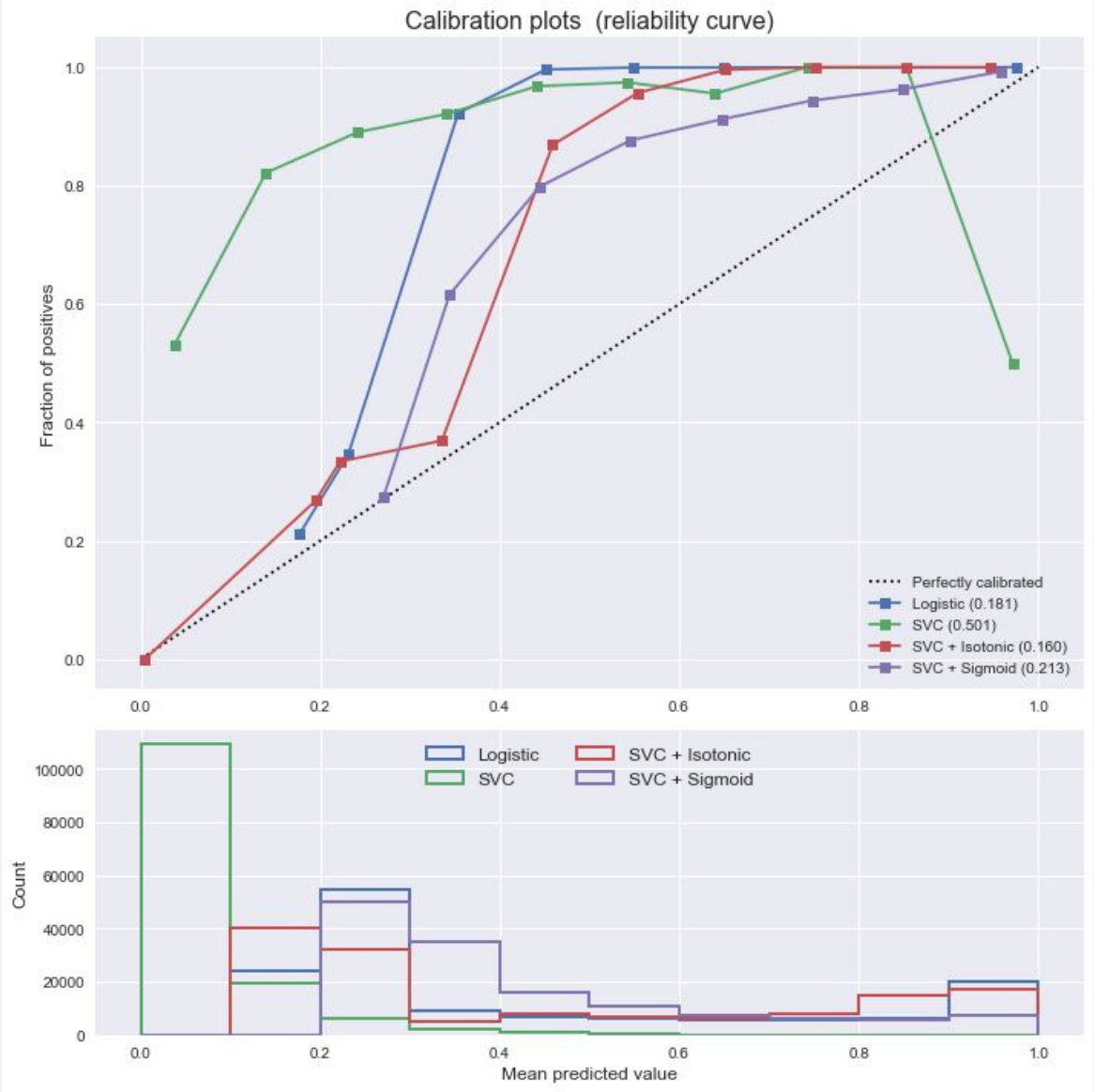
F1: 0.776

SVC + Sigmoid:

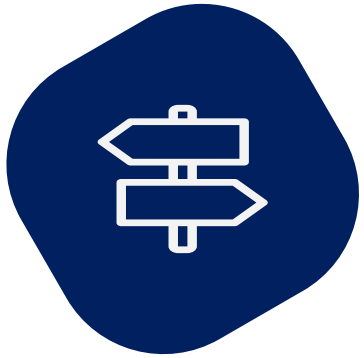
Precision: 0.931

Recall: 0.418

F1: 0.577



Reflections



- Deciding on features to create and developing them was the most challenging aspect.
- Given that the top F1 scores in the competition are 0.40XX, my model would likely score far less predicting the test orders.
- Additional questions to explore:
  - Which product is a customer likely to try for the first time during their next order?
  - When will a customer make their next order?
  - What customer segments can be derived from purchasing behavior?
  - What products are commonly purchased together?

Thank You

Sean Matthews

github/sean-io/market-basket-analysis